DOCUMENT RESUME

ED 423 278 TM 029 083

AUTHOR Kwak, Nohoon; Davenport, Ernest C., Jr.; Davison, Mark L.
TITLE A Comparative Study of Observed Score Approaches and Purification Procedures for Detecting Differential Item Functioning.
PUB DATE 1998-00-00
NOTE 60p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Diego, CA, April 12-16, 1998).
PUB TYPE Numerical/Quantitative Data (110) -- Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Ability; Comparative Analysis; Error of Measurement; *Estimation (Mathematics); *Item Bias; Sample Size; *Scores; *Tables (Data); *Test Items
IDENTIFIERS Item Bias Detection; *Mantel Haenszel Procedure; *Purification (Statistics)

ABSTRACT

The purposes of this study were to introduce the iterative purification procedure and to compare this with the two-step purification procedure, to compare false positive error rates and the power of five observed score approaches and to identify factors affecting power and false positive rates in each method. This study used 2,400 data sets that were divided into uniform, symmetric nonuniform, and nonsymmetric nonuniform differential item functioning (DIF) data sets. The sample size pairs were either 500,500 or 1,000,1,000 for the reference group and the focal group when the means of ability distributions for the 2 groups were the same, and either 1,000,500 or 1,000,250 for the reference and focal groups when the means of ability distributions for the 2 groups were different. Each dataset included four items with uniform, symmetric nonuniform, or nonsymmetric nonuniform DIF, with each DIF item having either a 0.4 or 0.8 amount of DIF (that is, the area between two item characteristic curves). The purification procedures reduced false positive error rates and/or increased power. The Mantel Haenszel method was superior to other methods with uniform DIF data sets, and the Absolute Mean Deviation method using the iterative purification procedure was superior to the others in nonuniform data sets when the means of ability distributions for the two groups were different. The ability estimation and the sample size affected detection rates and false positive error rates for all methods. The DIF effect size was also a strong influence on detection rates. (Contains 21 tables and 25 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 423 278

A COMPARATIVE STUDY OF OBSERVED SCORE APPROACHES AND PURIFICATION PROCEDURES FOR DETECTING DIFFERENTIAL ITEM FUNCTIONING

RUNNING HEAD: Iterative Purification Procedure for DIF

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Nohoon Kwak
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Nohoon Kwak
University of Minnesota

Ernest C. Davenport, Jr.
University of Minnesota

Mark L. Davison
University of Minnesota

Correspondence concerning this manuscript should be sent to Nohoon Kwak, Department of Educational Psychology, University of Minnesota, 178 Pillsbury Dr. S. E., Minneapolis, MN 55455

TM029083

ABSTRACT

The purposes of this study were to introduce the iterative purification procedure and to compare this with the two-step purification procedure, to compare false positive error rates and power of five observed score approaches, and to identify factors affecting power and false positive error rates in each method. This study used 2,400 data sets which were divided into uniform, symmetric nonuniform, and nonsymmetric nonuniform DIF data sets. The sample size pairs were either (500, 500) or (1,000, 1,000) for the reference group and the focal group when the means of ability distributions for the two groups were the same, and either (1,000, 500) or (1,000, 250) for the reference and the focal groups when the means of ability distributions for the two groups were different. Each dataset included 4 items with uniform, symmetric nonuniform, or nonsymmetric nonuniform DIF, with each DIF item having either a .4 or .8 amount of DIF (that is, the area between two item characteristic curves). The purification procedures reduced false positive error rates and/or increased power. The MH method was superior to other methods with uniform DIF data sets, and the AMD method using the iterative purification procedure was superior to the others in nonuniform data sets when the means of ability distributions for the two groups were different. The ability distribution and the sample size affected detection rates and false positive error rates for all methods. The DIF effect size was also a strong influence on detection rates.

Key words: iterative purification procedure, observed score approaches, differential item functioning (DIF), uniform DIF, nonuniform DIF

A Comparative Study of Observed Score Approaches and Purification Procedures for Detecting Differential Item Functioning

Nohoon Kwak, Ernest C. Davenport, Jr., and Mark L. Davison

University of Minnesota

Differential item functioning (DIF) has been an important issue in educational and psychological measurement in recent years. DIF exists if equally able individuals from different groups have unequal probabilities of answering an item correctly (Holland & Thayer, 1988; Shepard, Camilli, & Averill, 1981).

One major issue in DIF research is the purification of the matching variable. The rationale for purification procedure is that items with DIF will degrade ability estimation, which in turn may adversely affect the detection of DIF. Holland and Thayer (1988) suggested a two-step procedure to purify the matching variable by eliminating items with a preliminary indication of DIF. Several studies (Kwak, 1994; Miller & Oshima, 1992) indicated that the 2-step purification procedure had the positive effects of (a) reducing false positives for the Mantel-Haenszel (MH) and/or the full chi-square (FC) methods, and (b) improving power for the absolute mean deviation (AMD) method when the ability distributions for the two groups were approximately the same. However, Kwak, Davison, and Davenport (March, 1997) found that the two step purification procedure inflated false positive error rates and reduced power for the MH and the AMD methods when the distributions for the two groups were unequal. Therefore, another purification procedure

for DIF analysis is needed because the two-step purification procedure is effective only for the condition in which there are equal ability distributions for the two groups and equal ability distributions as an unresonable assumption for some situations.

Objective

The primary purpose of this study is to introduce a new purification procedure, the iterative purification procedure, and to compare this purification procedure with the two-step purification procedure suggested by Holland and Thayer (1988). The secondary purpose is to compare several observed score approaches based on their false positive error rates and power. The final purpose is to identify factors affecting power and false positive error rates in each observed score approach.

Methods

Data Generation

The current study used two thousand and four hundred simulated 40-item tests based on the three-parameter logistic model (Birnbaum, 1968). This model was selected because the three-parameter model reflects real data from standardized achievement tests (Ansley & Forsyth, 1985). The item parameters of 36 unbiased items were randomly chosen from parameter distributions of the ACT mathematics test (Drasgow, 1987). The means and the standard deviations were 1.09 and 0.35, 0.50 and 0.61, and 0.14 and 0.04 for the difficulty, b, the discrimination, a, and the pseudo-guessing, c, parameters for the ACT mathematics test, respectively.

Dichotomous item responses were generated by computing the probability of a correct response for each item and each examinee using the three-parameter logistic model.

The probability of a correct response was compared to a number sampled from the (0, 1) uniform distribution. If the probability was less than the random number, the item response was coded 0; otherwise, the item response was coded 1.

Ability Distribution

Shealy and Stout (1993) showed that the MH method yielded good adherence to the nominal significance levels even for differences in ability as large as one standard deviation, but Narayanan and Swaminathan (1994) argued that the difference in ability distribution inflated the Type I error rates on the MH method.

Hedges and Nowell (1995) reported that for gender differences in mathematics, the mean differences were minimal but the variance differences were 5 to 20%. To simulate a condition similar to that for gender differences in mathematics, this study used no difference in mean ability and a .2 difference in variance to create an equal mean ability condition.

Shepard, Camilli, and Williams (1984) reported that the difference between white and black seniors was .81σ with respect to the white's standard deviation for the math test in the High School and Beyond data. However, when they rescaled the combined samples to have a mean 0 and σ=1, the difference between white and black seniors was .7σ (Shepard, Camilli, & Williams, 1985). To simulate a condition similar to that for ethnic differences in achievement, this study used a .7 standard deviation difference in mean ability but no difference in variance between the reference and focal groups to create an unequal mean condition.

Sample Size

As sample size increases, the power for detecting DIF increases (Mazor, Clauser, &

Hambleton, 1992; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990). In comparisons of the equal mean ability condition, the sample size pairs were either (500, 500) or (1,000, 1,000) for the reference group and the focal group. In comparisons of unequal mean ability conditions, this study used sample size pairs of (1000, 500) or (1000, 250) for the reference and the focal groups.

Simulated DIF

There are two kinds of DIF, uniform DIF and nonuniform DIF (Mellenbergh, 1982). This study uses uniform and nonuniform DIF conditions because both appear in empirical studies (Bennett, Rock, & Kaplan, 1987; Ellis, 1989; Hambleton & Rogers, 1989; Linn, Levine, Hastings, & Wardrop, 1981; Mellenbergh, 1983). The nonuniform DIF condition includes 2 different DIF conditions: symmetric nonuniform DIF in which the *b* parameters for the two groups are the same but the *a* parameters for the two groups differ; nonsymmetric nonuniform DIF in which both the *a* and *b* parameters differ across groups.

Eight items with uniform DIF were obtained by varying the level of the *b* parameters (low, medium, or high) within the levels of the *a* parameter (low or high). For systematic nonuniform DIF, eight items were obtained by varying the level of the *a* parameters (low or medium) within the levels of the *b* parameter (low, medium or high). For nonsymmtric nonuniform DIF, eight items were obtained by varying both the level of the *b* parameter (low, medium, or high) and the level of the *a* parameters (low, medium or high). In each type of DIF, these eight items with DIF were assigned into two separate test. The first test included the low and high *b* parameter conditions. The second test included the medium *b* parameter conditions. The size of DIF was .4 or 8 in the amount of area between the two

item characteristic curves for reference and focal groups. One hundred replications were simulated for each type of DIF item (combination of difficulty and discrimination). Tables 1, 2, and 3 show the uniform, symmetric nonuniform, and nonsymmetric nonuniform DIF conditions.

Insert Tables 1, 2, and 3 Here

Test Statistics for Detecting DIF

This study used five test statistics; the Mantel-Haenszel (MH), the full chi-square (FC), the absolute mean deviation (AMD), the simultaneous item bias (SIB) test, and logistic regression (LR) methods; for the uniform DIF datasets but only four test statistics; the MH, the FC, the AMD, and the LR methods; for the nonuniform DIF datasets. These five test statistics were selected for this study because previous researches (Kwak, 1994; Kwak, Davenport, & Davison, 1997; Rogers & Swaminathan, 1993; Shealy & Stout, 1993; Swaminathan & Rogers, 1990) have shown that these methods are promising methods for detecting uniform and/or nonuniform DIF.

Purification Procedure

For the MH, the AMD, and the FC statistics, both the two-step and the iterative purification procedures were used to remove possible contamination from the matching variable while, for the LR method, only the iterative procedure was used. When a test includes one or more items with sizable amounts of DIF and the sample size is large, the two-step procedure for the LR method may produce too many false positives. On the other

hand, the SIB test uses a regression correction procedure to prevent inflation of the false positive error rate when there is an ability difference between groups. This regression correction procedure has a function similar to the purification procedure. Hence, the SIB test should not require any purification procedure.

Two-step Purification Procedure.

The MH, the AMD, and the FC methods were computed in two steps as proposed by Holland and Thayer (1988). First, score groups were obtained using total scores based on all items, and then the statistics were computed for all items. Those items for which the test statistic exceeded the critical value at α=.05 or α=.01 were identified and labeled as potentially displaying DIF. Next, total scores were reconstituted after eliminating items previously identified as DIF, and then the test statistics were calculated once again.

Iterative Purification Procedure.

1. The first step is to compute statistics for all items in the test. Then the item with the highest (significant) value is identified.

2. In the second step, the item identified as DIF in the first step is eliminated from the test, and the score groups are reformed using the total scores in the reduced (n-1)-item test. The test statistics for all n items of the test are computed again. If the first eliminated item is still significant in the second step, the next item with the highest (significant) value is eliminated including the previously eliminated item. Score groups are reformed using scores on the test of (n-2) items, and the test statistics for all n items of the test are computed once again. In this step, however, if the previously eliminated item is not significant in the next step, the item is included for forming total scores but the item with

the newly highest (significant) value is eliminated from the test. Step 2 is repeated until the iteration procedure meets the termination criteria. However, this iterative procedure may not converge. If this procedure does not converge even if it reaches a predetermined number of iterations, it uses an alternative iterative procedure.

Alternative Iterative Purification Procedure.

1. The item identified as DIF in the first step is eliminated from the test scores, and the score groups are reformed using the total scores in the reduced (n-1)-item test. The test statistics for all n items of the test are computed again.

2. In the next step the highest (significant) two values are identified. Then these items are eliminated from the test scores, and score groups are reformed using scores on the test of (n-2) items, and the test statistics for all n items of the test are computed once again. If the previously eliminated items are still significant in the next step, these items are also eliminated from the test score. However, if the previously eliminated items are not significant in the next step, the item is included for forming total scores. Steps are continued until the iteration procedure meets the termination criteria.

Termination Criteria.

The iterative procedure stops when it has iterated a predetermined number of iterations, when all the items in the reduced test have nonsignificant values, or when the previously identified items with DIF are the same as the present items with DIF.

Analysis

The statistical method used in the analysis of the false positive errors and power for the MH, the AMD, and the FC methods was a repeated measures ANOVA. For the LR and

the SIB procedures, it was a factorial ANOVA. However, the SIB test was excluded in the analysis for nonuniform DIF datasets since it is designed only for detecting uniform DIF. Although the MH method has the same function as the SIB test, it was included in the analysis for both uniform and nonuniform DIF datasets.

Because of the large sample sizes (i.e., 28,800 for the false positive error study; 3,200 for the power study), the results of the ANOVA would be statistically significant even if there were small mean differences. For meaningful interpretation, this study calculated the effect size (ES) proposed by Cohen (1988) to investigate the relative effect size for factors in the design. The index of the effect size used in the ANOVA design is nondirectional unlike the d index for the two sample test because the F ratio is nondirectional. Cohen (1988) suggested the following guidelines for interpreting effect size. He has defined a small effect as ES=.02, a medium effect as ES=.15, and a large effect as ES=.35 for a multivariate analysis. In the same context, effect sizes of .10, .25, and .40 are defined as small, medium, and large effects for an ANOVA analysis, respectively.

Results

Results of the study are presented in two parts. Part I contains the results for the false positive error rates and Part II contains the results of the relative power.

The false positive error study used a total of 28,800 items with no DIF which were divided into four conditions--that is, a combination of two ability distributions and two sample sizes. The power study used 3,200 items with DIF which were divided into eight conditions--that is, a combination of two ability distributions, two sample sizes, and two DIF sizes. In each study, uniform, symmetric nonuniform, and nonsymmetric nonuniform

DIF datasets were analyzed separately.

We interpreted the results for the interactions and then interpreted the results for the main effects or the simple effects. If there was no significant interaction, we discussed the results for the main effects. However, when there was a significant interaction, we interpreted a simple effect which was the effect of one factor at one level of the other factor. There are two kinds of interactions: ordinal and disordinal. A disordinal interaction occurs when the lines of the group means cross, and an ordinal interaction occurs when the lines do not cross but are nonparallel.

False Positive Error Study

In order to investigate false positive (FP) errors, a two between and one within factor repeated-measure design (Winer, 1962) was used for the MH, the AMD, and the FC methods while a two-way factorial ANOVA design was used for the LR and the SIB methods. The dependent variable was the FP error rate for both ANOVA designs. The two between factors were the ability distribution (i.e., same mean but different variance vs. different mean but same variance) and the sample size (i.e., large vs. small) for both designs. The one within factor was the purification procedure (i.e., two-step vs. iterative) for the repeated-measures design. Although the no-purification condition was not analyzed in the repeated-measures design, descriptive information was presented as a baseline against which to evaluate the effectiveness of the purification procedures.

Uniform DIF Datasets

The MANOVA and the ANOVA results are presented in Tables 4 and 5, respectively, and the corresponding percentages of false positives are shown in Table 6.

Because of the large sample size, there were statistically significant interactions and main effects.

Insert Tables 4, 5, and 6 Here

Tables 4 and 5 show that several significant interaction effects were observed but, for the most part, the effect sizes (ES) corresponding to interactions were fairly small (i.e., ES<.04 for MANOVA; ES<.10 for ANOVA). Interactions with noticeable effect size were those between the ability distribution and the purification procedure (ES=.16 at the α=.01 level; ES=.31 at the α=.05 level) for the AMD method and between the ability distribution and the sample size (ES=.23 at the α=.01 level) for the LR method. However, all the interactions were ordinal (see Table 6).

The purification procedure had a strong effect on the false positive error rates for the AMD method but not for the MH and the FC methods (see Table 4). Results in Table 6 indicate that for the AMD method, the two-step purification procedure produced much higher than expected false positive error rates for the ability distributions with same means but different variances at both α levels while the iterative purification procedure yielded approximately the expected and/or lower false positive error rates. Moreover, the AMD method using the 2-step purification procedure produced much higher false positive error rates than the AMD method using the iterative purification procedure when ability distributions for the two groups had equal means while the former produced similar false positive error rates to the latter when ability distributions for the two groups had different

means. Although there were no big differences between two-step and iterative purification procedures for the MH and the FC methods, the iterative procedure tends to slightly increase the false positive error rates in the different mean but same variance condition for the MH while it tends to decrease the false positives in the same condition for the FC method.

Table 4 shows that the ability distribution affected the false positive error rates for all the methods at both α levels. However, the effect sizes for the MH and the FC methods at the α=.01 level were .03 and .07, respectively. On the other hand, effect sizes of the ability distribution for all other methods were at least .10 at the α=.01 level. The sample size appeared to affect false positive error rates for all but one method but effect sizes were small (ES<.04) except for the LR method at both α levels. For the LR method, the false positive error rates were much higher than the nominal α levels, especially when the ability distributions for the two groups had different means and the sample size was large (see Table 6).

'Overall, the two-step procedure for the MH method produced fewer false positives in all conditions. In the distribution with the same mean but different variance, the two-step and the iterative procedure for the MH and the FC methods produced fewer false positives. However, in the distribution with different means but the same variance condition, the iterative procedure for the AMD method was a competitor for the MH method.

Symmetric Nonuniform DIF Datasets

Tables 7 and 8 show the MANOVA and ANOVA results, respectively, and Table 9 presents the corresponding percentage of false positives. The results had almost the same

pattern as those for uniform DIF datasets.

Insert Tables 7, 8, and 9 Here

There were several statistically significant interactions (see Tables 7 & 8). For the most part, effect sizes were less than .10. However, effect sizes of interactions between the ability distribution and the purification procedure (ES=.18 at the α=.01 level; ES=.33 at the α=.05 level) for the AMD method and between the ability distribution and the sample size (ES=.20 at the α=.01 level) for the LR method were noticeable. Fortunately, these interactions were ordinal (see Table 9).

Table 7 shows that the difference between the two purification procedures produced significant effects for the MH and the AMD methods at the α=.01 level and for all three methods at the α=.05 level. However, effect sizes were small for the MH method (i.e., ES=.03 at the α=.01 level) and for the FC (i.e., ES=.01 at the α=.05 level). For the AMD method, the two-step purification procedure yielded at least 5 times higher than expected false positive error rates at both α levels in the distributions with the same mean but different variance while the iterative purification produced approximately the expected or lower false positive error rates. Additionally, the AMD method using the 2-step purification procedure produced much higher false positive error rates than the AMD method using the iterative purification procedure for the ability distributions of the two groups with the same ability mean while the former produced similar false positive error rates to the latter for the ability distributions of the two groups with different ability means. For the MH and the FC

methods, there were no differences between the two purification procedures on the ability distributions with the same mean but different variances. However, the iterative purification produced fewer false positives than the two-step purification procedure in the ability distributions with different means but the same variance.

The ability distribution had a significant effect for all four methods at both α levels (see Tables 7 & 8). The MH, the FC, and the LR methods produced higher false positive error rates in the ability distributions with different means but the same variance. Conversely, the AMD method yielded higher false positive error rates in the ability distributions with the same means but different variance (see Table 9).

The sample size had statistically significant effects ($p<.001$) for the AMD and the LR methods at the $\alpha=.01$ level, and it had significant effects ($p<.01$) for the AMD, the FC, and the LR methods at the $\alpha=.05$ level. However, effect sizes were minimal for the AMD (i.e., ES=.02 at the $\alpha=.01$ level; ES=.03 at the $\alpha=.05$ level) and the FC method (ES=.05 at the $\alpha=.01$ level) while it was moderate for the LR method (ES=.23 at the $\alpha=.01$ level; ES=.27 at the $\alpha=.05$ level). For the LR method, the large sample sizes produced much higher false positive error rates than the nominal α levels.

Overall, the iterative procedure for MH method produced fewer false positives than the others. When ability distributions for two groups had the same mean but a minor difference in variance, the MH and the FC methods had fewer false positives. However, when ability distributions for two groups had different means but the same variance, the iterative procedure for the MH and the AMD produced fewer false positive error rates at both α levels.

Nonsymmetric Nonuniform DIF Dataset

The MANOVA and ANOVA results are presented in Tables 10 and 11, respectively, and the corresponding descriptive information is shown in Table 12. There were several significant interactions and main effects because of the large sample size.

Insert Tables 10, 11, and 12 Here

Table 10 shows that although many interactions were statistically significant ($p<.05$), for the most part, corresponding effect sizes were quite small ($ES<.05$) at both α levels. There were some considerable effect sizes including the interactions between the ability distribution and the purification procedure for the AMD at both α levels and between the ability distribution and the sample size for the LR method at the $\alpha=.01$ level. However, these interactions appeared to be ordinal (see Table 12).

Table 10 indicates that the purification effects were statistically significant ($p<.05$) for the AMD and the FC methods at the $\alpha=.01$ level and for all three methods at the $\alpha=.05$ level. For the MH and the FC methods, the purification effects were relatively small ($ES<.05$). However, for the AMD method, the purification procedure had a strong effect on the false positive error rates. For example, the iterative purification procedure reduced false positives by about 5% from those for no purification in the distributions with the same mean but different variances at the $\alpha=.01$ levels while the 2-step purification procedure increased false positive error rates at least 1.5% from those for no purification in the same condition (see Table 12). Another noticeable result was that the AMD method using the 2-step

purification procedure produced much higher false positive error rates than the AMD method using the iterative purification procedure for the ability distributions of the two groups with the same mean, while the former produced similar false positive error rates to the latter for the ability distributions of the two groups with different means.

The ability distribution effect was statistically significant ($p<.001$) for all four methods at both α levels. For the MH method, the difference between the two distributions was fairly small while for the other methods it was quite large, especially for the LR method. For the MH, the FC, and the LR methods, the false positive error rates were higher in the distributions with the different means but the same variance. However, for the AMD method, this difference was reversed (see Table 12).

The sample size effect was statistically significant ($p<.05$) for the MH, the FC, and the LR methods at the $\alpha=.01$ level and for the MH and the LR methods at the $\alpha=.05$ level. Excluding the LR method, the sample size effects for the other methods were minimal ($ES<.02$) at both α levels. The LR method yielded much higher than expected false positive error rates in the large sample ($ES>.25$).

Overall, the iterative procedure for the MH method produced fewer false positives than the others. In the distributions with the same mean but different variances, the MH and the FC methods had fewer false positives. However, in the distributions with different means but the same variance, the iterative procedure for the MH and the AMD produced fewer false positives than those of the other methods at both α levels.

Power Study

For the investigation of power, which is the ability to detect items with DIF, a three

between and one within factor repeated-measures design (Winer, 1962) was used for the MH, the AMD, and the FC methods while a three-way factorial ANOVA design was used for the LR and the SIB methods. The dependent variable was power for both designs. The within factor was the purification procedure and the three between factors were the DIF size (i.e., .4 of area vs. .8 of area); the ability distribution, and the sample size for the MH, the AMD, and the FC methods. For the LR and the SIB methods, the three independent variables were the same as the three between factors in the repeated-measures design. As in the false positive error study, the descriptive information for the no purification condition is presented as a baseline.

Uniform DIF

The MANOVA and ANOVA results are presented in Tables 13 and 14, respectively, and the corresponding detection rates are shown in Table 15. There were many statistically significant interaction effects ($p<.001$) because of the large sample size. The corresponding interaction effect sizes varied from small to medium. Fortunately, these interactions were ordinal and had relatively small effect sizes compared to effect sizes of main effects (see Table 15).

Insert Tables 13, 14, and 15 Here

Results in Table 15 show that for all five methods there were no differences in detection rates between the ability distributions with the same mean but different variances when the amount of DIF was .80 while there were large differences in detection rates

between the ability distributions with the same mean but different variances when the amount of DIF was .40.

Results in Table 13 indicate that the purification procedure was statistically significant for the AMD but not for the MH and the FC methods. For the AMD method, the iterative procedure improved detection rates (i.e., power) by as much as 50% from the result of no purification at the α=.01 level.

The ability distribution and the amount of DIF had strong effects on detection rates for all the methods. All the methods detected more items with DIF for the distributions with the same mean but different variances and in the .80 amount of DIF condition.

The sample size was statistically significant ($p<.001$) for the AMD, the FC, the LR, and the SIB methods. These methods produced more power in the large sample size condition.

Overall, the LR method detected more items with DIF in most conditions followed by the MH and the AMD methods. The LR method had more power for detecting items with DIF in the distributions with different means but the same variance while the MH, the AMD, and the LR methods had approximately the same power for detecting items with DIF in the distributions with the same mean but different variances.

Symmetric Nonuniform DIF

Tables 16 and 17 show the MANOVA and ANOVA results, respectively, and Table 18 presents the corresponding detection rates (i.e., power). There were several significant interaction effects ($p<.05$), but the corresponding effect sizes were generally small ($ES<.10$). Considerable two-way interactions were those between the ability distribution

and the amount of DIF for the MH, all two-way interactions for the AMD, the FC, and the LR methods, and between the sample size and the purification and between the amount of DIF and the purification procedure for the AMD method. There were also three-way interactions among between-factors for the AMD, the FC, and the LR methods. Unfortunately, three of the interactions were disordinal. These were interactions between the ability distribution and the purification procedure for the MH and between the ability distribution and the amount of DIF for the AMD and the FC methods. However, these disordinal interaction effects were minimal because the mean difference in one direction was fairly small while that in the other direction was large.

Insert Tables 16, 17, and 18 Here

Results in Table 16 indicate that the difference between the purification procedures has a significant effect for the MH and the AMD methods. Table 18 shows that for the MH and the AMD methods, detection rates were higher for the iterative procedure. For the AMD method, the iterative procedure produced much higher detection rates in all but one condition. For the MH method, the iterative purification procedure had higher detection rates for the distributions with different means but the same variance while detection rates for the two-step purification procedure were the same for the ability distributions with the same mean.

For all four methods, the effect of ability distribution was statistically significant ($p<.01$). The MH, the AMD, and the FC methods detected more items with DIF for the

ability distributions with different means but the same variance while the LR method identified more items with DIF for the ability distributions with the same mean but different variances.

The sample size and the amount of DIF had strong effects on detection rates for all four methods. All four methods had more power for detecting items with DIF with the large sample size and the .80 amount of DIF.

Overall, the LR method had the highest detection rates in most conditions followed by the AMD method. The LR method detected more items with DIF for the distributions with different means but the same variance and for the small amount of DIF (i.e., area=.40) for the distributions with the same mean but different variances.

Nonsymmetric Nonuniform DIF

Tables 19 and 20 show the MANOVA and the ANOVA results and Table 21 presents the corresponding power. There were many significant interactions ($P<.05$). All the significant interactions were ordinal at both α levels and the corresponding effect sizes were relatively small. Although for the AMD and the FC methods, interaction effect sizes between the ability distribution and the sample size at the α=05 level were noticeable (ES=.14 and .13 for the AMD and the FC, respectively), these were relatively small compared to effect sizes of main effects.

Insert Tables 19, 20, and 21 Here

Table 19 shows that the purification procedure had strong effects on detection rates for the MH and the AMD methods at the α=01 level and for all three methods at the α=01 level. For the MH and the FC methods, the iterative purification procedure identified more items with DIF only for the ability distributions with different means but the same variance while for the AMD method, the iterative procedure detected more items with DIF in most conditions (see Table 21).

For all four methods, all main effects were statistically significant ($p<.01$) at both α levels except the ability distribution effect for the FC methods at the α=05 level. The MH method detected more items with DIF for the ability distributions with different means but the same variance while the other three methods identified more items with DIF for the ability distributions with the same mean but different variances. On the other hand, all four methods detected more items with DIF with the large sample size and the large (i.e., area=.80) amount of DIF.

Overall, the LR method had the highest detection rates followed by the AMD method. The LR method detected more items with DIF in all but one small (i.e., area=.40) DIF size condition while the iterative procedure for the AMD method identified more items with DIF with the small sample size and for the ability distributions with different means but the same variance.

Discussion and Recommendations

For the uniform DIF datasets, the MH method appears to be the most effective method for detecting uniform DIF. It had fewer false positives when the ability distributions for two groups had the same mean. The AMD method using iterative purification was a

competitor to the MH method when the ability distributions for the two groups had different means.

Several researchers (Rogers & Swaminathan, 1993; Shealy & Stout, 1993; Swaminathan & Rogers, 1990) argued that the LR or the SIB method was as effective for detecting uniform DIF as the MH method. However, this study found different results. The LR method committed too many false positive errors with the large sample size or for ability distributions with different means but the same variance. The LR method had detection rates and false positive error rates similar to the MH only for limited conditions such as the small sample size with ability distributions having the same mean but different variances. The SIB method had problems similar to the LR method. The SIB method had much higher false positive error rates than the nominal α levels, particularly with the small sample size and the ability distributions with the same mean but different variances.

The purification procedure had an effect on detection rates for the MH and the AMD methods, and it affected false positive error rates for the AMD method. The ability distribution had a strong effect on both detection rates and false positives for all the methods. The sample size also affected detection rates for the AMD, the FC, the LR, and the SIB methods whereas it affected false positive error rates only for the LR and the SIB methods.

In detection of both symmetric and nonsymmetric nonuniform DIF, the AMD, the FC, and the LR methods were superior to the MH method. However, the MH method tends to produce fewer false positives in most conditions. Based on both detection rates and false positive error rates, the LR method at the α=.01 level or the FC method at the α=.05 level

appears to be the most effective method for detecting symmetric and nonsymmetric nonuniform DIF when the ability distributions for the two groups have the same mean, whereas the AMD method using iterative purification is the most effective method for detecting symmetric nonuniform DIF at both α levels when the ability distributions for the two groups have different means.

The purification procedure had an effect on detection rates and false positive error rates for the MH and the AMD methods, but the effect of the purification procedure on false positive error rates for the MH method was minimal. The ability distribution was a strong factor for both detection rates and false positive error rates for all four methods. The sample size and the DIF size were two strong factors for detection rates for all four methods, and the sample size also affected false positive error rates for the AMD and the LR methods.

Overall, when a test contains some items with DIF, the purification procedure leads to fewer false positives and to higher detection rates. However, when the ability distributions for the two groups have the same mean, the two-step purification procedure for the AMD was a marked exception. It tends to inflate false positive error rates when ability distributions for two groups have the same mean but different variances.

The results with respect to ability distribution were consistent with the results reported by Shealy and Stout (1993). The MH and the other methods did not inflate the Type I error rates even though there were differences in ability as much as .7 standard deviations. Moreover, all methods had higher detection rates for the ability distributions with the same mean but different variances as compared to the condition with ability

distributions having different means but the same variance. However, this result may be caused by the ability distribution used in this study. The ability distributions used in this study were similar to those used by Shealy and Stout (1993), and these were narrower than those used by Kwak and his colleagues (Kwak, Davison, & Davenport, March, 1997) and Narayanan and Swaminathan (1994).

The results related to sample size were consistent with the results reported by Mazor, Clauser and Hambleton (1992), Rogers and Swaminathan (1993), and Swaminathan and Rogers (1990). As sample size increases, the detection rate also increases. However, the sample size functions differently for the false positive error rate with each method. As the sample size decreases the false positive error rate increases for the MH, the FC, and the SIB methods but it decreases for the AMD and the LR methods.

Two major recommendations arise from this study. First, although both purification procedures (except the two-step for the AMD method) lead to false positive error rates close to or less than the nominal α levels and higher detection rates, the iterative purification procedure surpasses the two-step purification procedure in detection rates and false positive error rates. Second, those using an observed score approach should use the Mantel-Haenszel method with a nonuniform detection method such as the absolute mean deviation, the full chi-square, or the logistic regression statistic. The choice depends on the combination of the ability distribution and the sample size. When the means of ability distributions for the two groups are approximately the same and one of the two groups has small sample size (i.e., less than 500), those using observed score approaches may want to combine the MH with the LR method. When the means of ability distributions for the two

groups are different or the two groups have large sample sizes (i.e., larger than 500), the MH method with the AMD method using iterative purification procedure is recommended.

References

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristic of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9*, 37-48.

Bennet, R. E., Rock, D. A., & Kaplan, B. A. (1987). SAT differential item performance for nine handicapped groups. *Journal of Educational Measurement, 24*, 56-64.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* NJ: Erlbaum.

Drasgow, F. (1987). A study of measurement bias of two standard psychological tests. *Journal of Applied Psychology, 72*, 19-30.

Ellis, B. (1989). Differential item functioning: Implications for test translations. *Journal of Applied Psychology, 74*, 912-921.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2*, 313-334.

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science, 269*, 41-45.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the

Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kwak, N. (1994). *A simulation study of methods for detecting uniform and nonuniform item bias*. Unpublished master thesis, University of Minnesota.

Kwak, N., Davenport, Jr. E. C., & Davison, M. L. (1997). *A comparative study of the Mantel-Haenszel and full chi-square methods for detecting uniform and nonuniform DIF*. Manuscript for Publication.

Kwak, N., Davison, M. L., & Davenport, Jr. E. C. (1997, March). *An unsigned Mantel-Haenszel statistic for detecting uniform and nonuniform DIF*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). An investigation of item bias in a test of reading comprehension. *Applied Psychological Measurement, 5*, 159-173.

Mazor, K. M., Clauser, B., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52*, 443-451.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Measurement, 7*, 105-118.

Mellenbergh, G. J. (1983). Conditional item bias methods. In S. H. Irvine & J. W. Berry (Eds.), *Human assessment and cultural factors* (pp. 293-302). NY: Plenum Press.

Miller, M. D., & Oshima, T. C. (1992). Effect of sample size, number of biased items, and magnitude of bias on a 2-stage item bias estimation method. *Applied*

Psychological Measurement, 16, 381-388.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18*, 315-328.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*, 105-116.

Shealy, R., & Stout, W. (1993). A model based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Shepard, L. A., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6*, 317-375.

Shepard, L. A., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics, 9*, 93-128.

Shepard, L. A., Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement, 22*, 77-105.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.

Winer, B. J. (1971). *Statistical principles in experimental design*. New York: McGraw-Hill.

Table 1

Item Parameters Used to Generate Items with Uniform DIF

Item Type	DIF Size	a_R	b_R	a_F	b_F
Low *b*, High *a*					
	.4	1.44	-0.42	1.44	0.08
	.8	1.44	-0.78	1.44	0.24
Medium *b*, Low *a*					
	.4	0.74	0.16	0.74	0.73
	.8	0.74	-0.05	0.74	1.10
Medium *b*, High *a*					
	.4	1.44	0.26	1.44	0.74
	.8	1.44	0.05	1.44	1.01
High *b*, Low *a*					
	.4	0.74	0.82	0.74	1.42
	.8	0.74	0.66	0.74	1.92

Table 2

Item Parameters Used to Generate Items with Symmetric Nonuniform DIF

Item Type	DIF Size	a_R	b_R	a_F	b_F
Low *b*, Medium *a*					
	.4	0.64	0.00	1.44	0.00
	.8	0.32	0.00	1.84	0.00
Medium *b*, Low *a*					
	.4	0.35	0.30	0.81	0.30
	.8	0.11	0.30	1.04	0.30
Medium *b*, Medium *a*					
	.4	0.65	0.30	1.44	0.30
	.8	0.31	0.30	1.74	0.30
High *b*, Medium *a*					
	.4	0.64	1.00	1.44	1.00
	.8	0.32	1.00	1.84	1.00

Table 3

Item Parameters Used to Generate Items with Nonsymmetric Nonuniform DIF

Item Type	DIF Size	a_R	b_R	a_F	b_F
Low *b*, Medium *a*					
	.4	0.66	-0.20	1.54	0.00
	.8	0.29	-0.20	1.74	0.00
Medium *b*, Low *a*					
	.4	0.33	0.20	0.78	0.40
	.8	0.12	0.20	1.05	0.40
Medium *b*, High *a*					
	.4	0.70	0.20	1.51	0.40
	.8	0.33	0.20	1.82	0.40
High *b*, Low *a*					
	.4	0.35	1.00	0.78	1.20
	.8	0.12	0.95	0.98	1.25

Table 4

Repeated-Measures Multivariate Analysis of Variance Comparing Purification Procedures and Effects of All Factors for the Mantel-Haenszel, the Absolute Mean Deviation, and the Full Chi-square Methods on the False positive Error Rate in the Uniform DIF Datasets

(df=1, 28796)

	MH		AMD		FC	
$\alpha = .01$	F	ES	F	ES	F	ES
			Between Factors			
Ability (D)	26.22***	.03	721.18***	.16	144.89***	.07
Sample (S)	9.16**	.02	18.13***	.03	5.80*	.01
D × S	9.16**	.02	29.09***	.03	5.80*	.01
			Within Factors			
Purif. (P)	10.29***	.02	822.14***	.17	20.91***	.03
D × P	10.29***	.02	718.18***	.16	20.91***	.03
S × P	2.57	.01	4.25*	.01	2.98	.01
D × S × P	2.57	.01	3.75	.01	2.98	.01
$\alpha = .05$	F	ES	F	ES	F	ES
			Between Factors			
Ability (D)	228.23***	.09	3547.19***	.35	991.32***	.19
Sample (S)	10.35***	.02	20.26***	.03	.29	.00
D × S	10.35***	.02	24.96***	.03	.29	.00
			Within Factors			
Purif. (P)	86.01***	.05	3370.98***	.34	35.44***	.04
D × P	86.01***	.05	2718.01***	.31	35.44***	.04
S × P	8.05**	.02	47.31***	.04	.10	.00
D × S × P	8.50**	.02	56.80***	.04	.10	.00

*P<.05. **P<.01. ***P<.001.

Table 5

Analysis of Variance Comparing Effects of All Factors on the False Positive Error Rate for the Logistic Regression and the SIB Procedures in the Uniform DIF Datasets (df=1, 28796)

	LR		SIB	
	F	ES	F	ES
$\alpha = .01$				
Ability Dist. (D)	3425.97***	.34	287.41***	.10
Sample Size (S)	1942.72***	.26	170.80***	.08
D × S	1534.43***	.23	114.96***	.06
$\alpha = .05$				
Ability Dist. (D)	8338.30***	.54	209.85***	.09
Sample Size (S)	1384.31***	.22	20.08***	.03
D × S	25.40***	.03	54.37***	.04

***P<.001.

Table 6

Percentage of False positives in Each Condition for Each Method in the Uniform DIF Datasets

	Same Mean		Different Mean	
	Large Sample	Small Sample	Large Sample	Small Sample
$\alpha = .01$	FP (%)	FP (%)	FP (%)	FP (%)
MH				
No Purification	.1%	.0%	.1%	.3%
2-step	0%	0%	.0%	.2%
Iterative	0%	0%	.1%	.3%
AMD				
No Purification	.8%	1.4%	.4%	.6%
2-step	7.6%	6.0%	.5%	.6%
Iterative	1.0%	.3%	.3%	.4%
FC				
No Purification	.0%	.1%	.9%	1.0%
2-step	0%	0%	.7%	1.2%
Iterative	0%	0%	.5%	.7%
LR	1.7%	.1%	33.3%	6.4%
SIB	1.6%	5.7%	.6%	1.0%

Continued next page

Table 6 continued

	Same Mean		Different Mean	
	Large Sample	Small Sample	Large Sample	Small Sample
$\alpha = .05$	FP (%)	FP (%)	FP (%)	FP (%)
MH				
No Purification	2.8%	.1%	1.8%	2.5%
2-step	0%	0%	.8%	1.1%
Iterative	0%	0%	1.3%	2.0%
AMD				
No Purification	4.9%	7.4%	2.3%	2.4%
2-step	29.6%	24.5%	2.3%	2.5%
Iterative	4.7%	5.3%	1.2%	1.2%
FC				
No Purification	.4%	.3%	4.8%	4.8%
2-step	0%	0%	4.8%	4.9%
Iterative	0%	0%	3.6%	3.8%
LR	20.0%	4.4%	66.6%	46.2%
SIB	10.5%	7.1%	4.1%	5.0%

Table 7

Repeated-Measures Multivariate Analysis of Variance Comparing Purification Procedures and Effects of All Factors for the Mantel-Haenszel, the Absolute Mean Deviation, and the Full Chi-square Methods on the False positive Error Rate in the Symmetric Nonuniform DIF Datasets (df=1, 28796)

	MH		AMD		FC	
$\alpha = .01$	F	ES	F	ES	F	ES
			Between Factors			
Ability (D)	76.87***	.05	791.87***	.17	181.95***	.08
Sample (S)	3.29	.01	11.43***	.02	.53	.00
D × S	3.29	.01	17.63***	.02	.53	.00
			Within Factors			
Purif. (P)	34.67***	.03	938.77***	.18	.95	.00
D × P	34.67***	.03	928.05***	.18	.95	.00
S × P	7.16**	.02	4.68*	.01	.64	.00
D × S × P	7.16**	.02	4.94*	.01	.64	.00
$\alpha = .05$	F	ES	F	ES	F	ES
			Between Factors			
Ability (D)	536.68***	.14	3793.96***	.36	1101.46***	.20
Sample (S)	2.13	.00	20.54***	.03	6.04*	.01
D × S	2.13	.00	22.49***	.03	5.37*	.01
			Within Factors			
Purif. (P)	223.01***	.09	3707.05***	.36	6.24*	.01
D × P	223.01***	.09	3061.65***	.33	6.24*	.01
S × P	3.51	.01	5.09*	.01	3.63	.01
D × S × P	3.51	.01	7.25**	.02	3.63	.01

*P<.05. **P<.01. ***P<.001.

Table 8

Analysis of Variance Comparing Effects of All Factors on the False Positive Error Rate for the Logistic Regression Procedure in the Symmetric Nonuniform DIF Datasets

(df=1, 28796)

	LR	
	F	ES
$\alpha = .01$		
Ability Distribution (D)	3289.73***	.34
Sample Size (S)	1514.14***	.23
D × S	1209.65***	.20
$\alpha = .05$		
Ability Distribution (D)	7354.77***	.51
Sample Size (S)	2047.11***	.27
D × S	150.40***	.07

***P<.001.

Table 9

Percentage of False positives in Each Condition for Each Method in the Symmetric Nonuniform DIF Datasets

	Same Mean		Different Mean	
	Large Sample	Small Sample	Large Sample	Small Sample
$\alpha = .01$	FP (%)	FP (%)	FP (%)	FP (%)
MH				
No Purification	0%	0%	.7%	.6%
2-step	0%	0%	.5%	.6%
Iterative	0%	0%	.1%	.4%
AMD				
No Purification	4.6%	4.3%	.6%	.6%
2-step	8.5%	7.0%	.5%	.6%
Iterative	.7%	.3%	.5%	.6%
FC				
No Purification	0%	0%	1.3%	1.4%
2-step	0%	0%	.9%	1.1%
Iterative	0%	0%	.9%	.9%
LR	1.6%	.3%	31.7%	7.7%

Continued next page

Table 9 continued

	Same Mean		Different Mean	
	Large Sample	Small Sample	Large Sample	Small Sample
$\alpha = .05$	FP (%)	FP (%)	FP (%)	FP (%)
MH				
No Purification	0%	0%	3.7%	3.7%
2-step	0%	0%	3.8%	3.9%
Iterative	0%	0%	1.7%	2.3%
AMD				
No Purification	17.1%	19.3%	2.7%	2.9%
2-step	29.7%	26.5%	2.3%	2.5%
Iterative	4.7%	3.5%	1.3%	1.2%
FC				
No Purification	.0%	.0%	5.7%	5.9%
2-step	0%	.0%	4.4%	5.5%
Iterative	0%	.0%	4.3%	4.6%
LR	21.0%	5.1%	68.4%	40.6%

Table 10

Repeated-Measures Multivariate Analysis of Variance Comparing Purification Procedures and Effects of All Factors for the Mantel-Haenszel, the Absolute Mean Deviation, and the Full Chi-square Methods on the False positive Error Rate in the Nonsymmetric Nonuniform DIF Datasets (df=1, 28796)

	MH		AMD		FC	
$\alpha = .01$	F	ES	F	ES	F	ES
			Between Factors			
Ability (D)	43.79***	.04	780.42***	.16	203.92***	.08
Sample (S)	7.64*	.02	1.19	.00	6.46*	.01
D × S	7.64*	.02	5.76*	.01	6.46*	.01
			Within Factors			
Purif. (P)	.07	.00	787.08***	.17	5.10*	.01
D × P	.07	.00	780.61***	.16	5.10*	.01
S × P	1.67	.00	28.20***	.03	1.18	.00
D × S × P	1.67	.00	30.71***	.03	1.18	.00
$\alpha = .05$	F	ES	F	ES	F	ES
			Between Factors			
Ability (D)	411.89***	.12	3912.17***	.37	1084.44***	.19
Sample (S)	6.61*	.02	2.71	.01	3.47	.01
D × S	6.61*	.02	12.06***	.02	2.78	.01
			Within Factors			
Purif. (P)	52.52***	.04	3268.01***	.34	17.99***	.02
D × P	52.52***	.04	2762.00***	.31	17.99***	.02
S × P	7.83**	.02	12.37***	.02	.28	.00
D × S × P	7.83**	.02	19.47***	.03	.28	.00

*P<.05. **P<.01. ***P<.001.

Table 11

Analysis of Variance Comparing Effects of All Factors on the False Positive Error Rate for the Logistic Regression Procedure in the Nonsymmetric Nonuniform DIF Datasets

(df=1, 28796)

	LR	
	F	ES
$\alpha = .01$		
Ability Distribution (D)	2803.07***	.31
Sample Size (S)	2161.50***	.27
D × S	1726.39***	.24
$\alpha = .05$		
Ability Distribution (D)	8022.68***	.53
Sample Size (S)	1787.52***	.25
D × S	79.05***	.05

***P<.001.

Table 12

Percentage of False positives in Each Condition for Each Method in the Nonsymmetric Nonuniform DIF Datasets

	Same Mean		Different Mean	
	Large Sample	Small Sample	Large Sample	Small Sample
$\alpha = .01$	FP (%)	FP (%)	FP (%)	FP (%)
MH				
No Purification	0%	0%	.2%	.4%
2-step	0%	0%	.2%	.4%
Iterative	0%	0%	.1%	.4%
AMD				
No Purification	4.9%	5.1%	.6%	.7%
2-step	8.4%	6.6%	.5%	.7%
Iterative	.3%	1.2%	.5%	.6%
FC				
No Purification	0%	0%	1.1%	1.3%
2-step	0%	0%	.9%	1.4%
Iterative	0%	0%	.8%	1.1%
LR	1.7%	.2%	30.4%	3.7%

Continued next page

Table 12 continued

	Same Mean		Different Mean	
	Large Sample	Small Sample	Large Sample	Small Sample
$\alpha = .05$	FP (%)	FP (%)	FP (%)	FP (%)
MH				
No Purification	0%	0%	2.3%	2.8%
2-step	0%	0%	2.6%	2.9%
Iterative	0%	0%	1.6%	2.5%
AMD				
No Purification	18.5%	21.4%	2.9%	2.8%
2-step	29.2%	26.3%	2.0%	2.6%
Iterative	4.6%	5.0%	1.2%	1.4%
FC				
No Purification	.0%	.0%	5.2%	5.9%
2-step	0%	.0&	4.8%	5.4%
Iterative	0%	.0%	4.0%	4.4%
LR	18.3%	2.5%	64.8%	40.6%

Table 13

Repeated-Measures Multivariate Analysis of Variance Comparing Purification Procedures and Effects of All Factors for the Mantel-Haenszel, the Absolute Mean Deviation, and the Full Chi-square Methods on the Detection Rate (Power) in the Uniform DIF Datasets (df=1, 3192)

	MH		AMD		FC	
$\alpha = .01$	F	ES	F	ES	F	ES
			Between Factors			
Ability (D)	18094.73***	2.38	7389.12***	1.52	2203.86***	.83
Sample (S)	1.93	.02	150.80***	.22	167.25***	.23
DIF Size (DS)	18001.33***	2.37	10630.30***	1.82	4812.40***	1.23
D × S	.12	.00	46.36***	.12	59.28***	.14
D × DS	17722.61***	2.36	6188.17***	1.39	1605.44***	.71
S × DS	.00	.00	32.65***	.10	50.43***	.13
D × S × DS	1.09	.02	178.87***	.24	183.07***	.24
			Within Factors			
Purificatn (P)	-	-	348.48***	.33	5.79*	.04
D × P	-	-	65.04***	.14	9.56**	.05
S × P	-	-	188.17***	.24	.00	.00
DS × P	-	-	112.43***	.19	.12	.00
D × S × P	-	-	49.27***	.12	.47	.01
D × DS × P	-	-	260.14***	.29	1.06	.02
S × DS × P	-	-	43.18***	.12	1.89	.02
D × S × DS × P	-	-	201.28***	.25	.47	.01

Continued next page

Table 13 continued

	MH		AMD		FC	
$\alpha = .05$	F	ES	F	ES	F	ES
			Between Factors			
Ability (D)	5444.37***	1.31	7093.78***	1.49	2748.70***	.93
Sample (S)	1.91	.02	76.49***	.15	70.23***	.15
DIF Size (DS)	5444.37***	1.31	9581.01***	1.73	4648.17***	1.21
D × S	1.91	.02	48.29***	.12	60.69***	.14
D × DS	5444.37***	1.31	6481.95***	1.43	2515.43***	.89
S × DS	1.91	.02	30.37***	.10	41.51***	.11
D × S × DS	1.91	.02	103.71***	.18	94.63***	.17
			Within Factors			
Purificatn (P)	10.40***	.06	192.48***	.25	.22	.00
D × P	10.40***	.06	43.54***	.12	.01	.00
S × P	.29	.00	120.94***	.19	.01	.00
DS × P	10.40***	.06	93.01***	.17	.43	.01
D × S × P	.29	.00	71.56***	.15	.01	.00
D × DS × P	10.40***	.06	117.25***	.19	.08	.00
S × DS × P	.29	.00	55.42***	.13	.01	.00
D × S × DS × P	.29	.00	144.30***	.21	.01	.00

*P<.05. **p<.01. ***P<.001.

Table 14

Analysis of Variance Comparing Effects of All Factors on the Detection Rate (Power) for the Logistic Regression and the SIB Procedures in the Uniform DIF Datasets (df=1, 3192)

	LR		SIB	
	F	ES	F	ES
$\alpha = .01$				
Ability Dist. (D)	1123.75***	.59	4860.85***	1.17
Sample Size (S)	312.63***	.31	86.43***	.16
DIF Size (DS)	1256.17***	.63	5776.70***	1.35
D × S	248.41***	.28	72.76***	.15
D × DS	1123.75***	.59	5404.11***	1.30
S × DS	312.63***	.31	1.11	.02
D × S × DS	248.41***	.28	4.86*	.04
$\alpha = .05$				
Ability Dist. (D)	154.99***	.22	2523.64***	.89
Sample Size (S)	63.36***	.14	56.54***	.13
DIF Size (DS)	154.99***	.22	3335.67***	1.02
D × S	63.36***	.14	31.35***	.10
D × DS	154.99***	.22	3225.71***	1.01
S × DS	63.36***	.14	2.56	.03
D × S × DS	63.36***	.14	.00	.00

*p<.05. ***P<.001.

Table 15

Detection Rate (Power) in Each Condition for Each Method in the Uniform DIF Datasets

	Same Mean				Different Mean			
	Large Sample		Small Sample		Large Sample		Small Sample	
$\alpha = .01$	.80	.40	.80	.40	.80	.40	.80	.40
MH								
No Purification	1.0	1.0	1.0	.97	1.0	.02	.99	.01
2-step	1.0	1.0	1.0	.99	1.0	.04	.99	.03
Iterative	1.0	1.0	1.0	.99	1.0	.04	.99	.03
AMD								
No Purification	1.0	.78	1.0	.48	.98	.01	.83	.01
2-step	1.0	.95	1.0	.58	.99	.01	.87	.03
Iterative	1.0	1.0	1.0	.98	1.0	.04	.99	.04
FC								
No Purification	1.0	.81	1.0	.51	.98	.01	.88	.02
2-step	1.0	.93	1.0	.58	.99	.03	.90	.04
Iterative	1.0	.93	1.0	.59	.99	.02	.89	.03
LR	1.0	1.0	1.0	.97	1.0	.72	1.0	.20
SIB	.99	1.0	.90	.86	1.0	.02	.99	.02

Continued next page

Table 15 continued

	Same Mean				Different Mean			
	Large Sample		Small Sample		Large Sample		Small Sample	
$\alpha = .05$	.80	.40	.80	.40	.80	.40	.80	.40
MH								
No Purification	1.0	1.0	1.0	1.0	1.0	.05	1.0	.06
2-step	1.0	1.0	1.0	1.0	1.0	.14	1.0	.11
Iterative	1.0	1.0	1.0	1.0	1.0	.16	1.0	.12
AMD								
No Purification	1.0	.90	1.0	.55	.99	.03	.91	.03
2-step	1.0	.99	1.0	.68	1.0	.03	.94	.05
Iterative	1.0	1.0	1.0	.98	1.0	.06	.99	.07
FC								
No Purification	1.0	.95	1.0	.65	1.0	.07	.95	.08
2-step	1.0	.99	1.0	.75	1.0	.09	.97	.11
Iterative	1.0	.99	1.0	.75	1.0	.10	.97	.12
LR	1.0	1.0	1.0	1.0	1.0	.95	1.0	.75
SIB	.99	1.0	.90	.88	1.0	.12	1.0	.09

Table 16

Repeated-Measures Multivariate Analysis of Variance Comparing Purification Procedures and Effects of All Factors for the Mantel-Haenszel, the Absolute Mean Deviation, and the Full Chi-square Methods on the Detection Rate (Power) in the Symmetric Nonuniform DIF Datasets (df=1, 3192)

	MH		AMD		FC	
α = .01	F	ES	F	ES	F	ES
			Between Factors			
Ability (D)	196.44***	.25	85.89***	.16	23.94***	.09
Sample (S)	145.69***	.21	331.12***	.32	335.07***	.33
DIF Size (DS)	153.46***	.22	1744.09***	.74	1598.16***	.71
D × S	10.60***	.06	14.34***	.07	20.51***	.08
D × DS	83.39***	.16	183.95***	.24	68.50***	.15
S × DS	3.48	.03	68.07***	.15	109.47***	.19
D × S × DS	4.10*	.04	28.82***	.10	44.39***	.12
			Within Factors			
Purif. (P)	52.77***	.13	579.31***	.43	.20	.00
D × P	57.79***	.13	9.83**	.06	1.42	.02
S × P	7.31**	.05	26.60***	.09	.80	.02
DS × P	.46	.01	102.27***	.21	.09	.00
D × S × P	9.25**	.05	.04	.00	.20	.00
D × DS × P	5.59*	.04	2.96	.03	.20	.00
S × DS × P	3.45	.03	22.59***	.08	2.69	.03
D × S × DS × P	4.82*	.04	8.02**	.05	.36	.01

Continued next page

Table 16 continued

	MH		AMD		FC	
$\alpha = .05$	F	ES	F	ES	F	ES
			Between Factors			
Ability (D)	131.68***	.20	85.73***	.16	24.86***	.09
Sample (S)	101.67***	.18	223.98***	.26	303.02***	.31
DIF Size (DS)	234.54***	.27	1156.58***	.60	1076.79***	.58
D × S	11.68***	.06	41.78***	.11	52.64***	.13
D × DS	76.27***	.15	86.81***	.16	38.20***	.11
S × DS	2.31	.03	68.38***	.15	214.93***	.26
D × S × DS	1.82	.02	36.68***	.11	69.40***	.15
			Within Factors			
Purif. (P)	13.27***	.06	475.94***	.39	.84	.02
D × P	10.75***	.06	.01	.00	.60	.01
S × P	20.73***	.08	55.05***	.13	.00	.00
DS × P	17.55***	.07	188.57***	.24	.12	.00
D × S × P	14.63***	.07	7.66**	.05	1.12	.02
D × DS × P	14.63***	.07	.20	.00	.60	.01
S × DS × P	4.78*	.04	.05	.00	.04	.00
D × S × DS × P	2.12	.03	4.91*	.04	1.79	.02

*$\underline{P}$<.05. **$\underline{P}$<.01. ***$\underline{P}$<.001.

Table 17

Analysis of Variance Comparing Effects of All Factors on the Detection Rate (Power) for the Logistic Regression Procedure in the Symmetric Nonuniform DIF Datasets (df=1, 3192)

	LR	
	F	ES
$\alpha = .01$		
Ability Distribution (D)	7.39**	.05
Sample Size (S)	169.78***	.23
DIF Size (DS)	350.83***	.33
D × S	6.65**	.05
D × DS	6.65**	.05
S × DS	145.92***	.21
D × S × DS	7.37**	.05
$\alpha = .05$		
Ability Distribution (D)	42.45***	.12
Sample Size (S)	83.85***	.16
DIF Size (DS)	174.43***	.23
D × S	42.45***	.12
D × DS	35.85***	.11
S × DS	74.45***	.15
D × S × DS	35.85***	.11

**P<.01. ***P<.001.

Table 18

Detection Rate (Power) in Each Condition for Each Method in the Symmetric Nonuniform DIF Datasets

	Same Mean				Different Mean			
	Large Sample		Small Sample		Large Sample		Small Sample	
$\alpha = .01$	.80	.40	.80	.40	.80	.40	.80	.40
MH								
No Purification	.69	.46	.53	.17	.79	.73	.69	.60
2-step	.71	.44	.53	.14	.77	.69	.59	.54
Iterative	.72	.43	.53	.13	.77	.75	.67	.62
AMD								
No Purification	1.0	.45	.86	.08	.97	.57	.80	.41
2-step	1.0	.44	.83	.06	.95	.58	.77	.37
Iterative	1.0	.64	.99	.23	1.0	.80	.91	.63
FC								
No Purification	1.0	.63	.93	.20	.99	.64	.86	.46
2-step	1.0	.62	.92	.14	.99	.65	.86	.44
Iterative	1.0	.64	.93	.14	.98	.66	.86	.43
LR	1.0	.99	.99	.71	1.0	.89	.99	.71

Continued next page

Table 18 continued

	Same Mean				Different Mean			
	Large Sample		Small Sample		Large Sample		Small Sample	
$\alpha = .05$	.80	.40	.80	.40	.80	.40	.80	.40
MH								
No Purification	.90	.50	.65	.39	.90	.76	.81	.70
2-step	.90	.50	.66	.34	.89	.75	.75	.64
Iterative	.90	.50	.66	.35	.85	.79	.79	.71
AMD								
No Purification	1.0	.59	.93	.18	1.0	.68	.91	.53
2-step	1.0	.62	.88	.17	.99	.66	.89	.53
Iterative	1.0	.76	.97	.44	1.0	.85	.97	.75
FC								
No Purification	1.0	.77	.99	.37	1.0	.79	.97	.64
2-step	1.0	.83	.98	.34	1.0	.79	.96	.60
Iterative	1.0	.84	.99	.33	1.0	.79	.96	.63
LR	1.0	.97	1.0	.93	1.0	.97	.99	.75

Table 19

Repeated-Measures Multivariate Analysis of Variance Comparing Purification Procedures and Effects of All Factors for the Mantel-Haenszel, the Absolute Mean Deviation, and the Full Chi-square Methods on the Detection Rate (Power) in the Nonsymmetric Nonuniform DIF Datasets (df=1, 3192)

	MH		AMD		FC	
$\alpha = .01$	F	ES	F	ES	F	ES
			Between Factors			
Ability (D)	146.34***	.21	12.42***	.06	7.63**	.05
Sample (S)	54.55***	.13	346.47***	.33	349.91***	.33
DIF Size (DS)	382.70***	.35	2380.63***	.86	2424.57***	.87
D × S	32.59***	.10	26.34***	.09	24.50***	.09
D × DS	6.94**	.05	.46	.01	4.56*	.04
S × DS	.01	.00	63.64***	.14	79.39***	.16
D × S × DS	1.63	.02	48.83***	.12	57.08***	.13
			Within Factors			
Purif. (P)	138.79***	.21	1175.31***	.61	3.07	.03
D × P	138.79***	.21	12.92***	.06	2.06	.03
S × P	2.83	.03	58.19***	.14	.63	.01
DS × P	3.37	.03	268.83***	.29	.03	.00
D × S × P	3.96*	.04	3.39	.03	.63	.01
D × DS × P	3.37	.03	63.66***	.14	.63	.01
S × DS × P	.00	.00	34.51***	.10	5.71*	.04
D × S × DS × P	.09	.00	.62	.01	2.06	.03

Continued next page

Table 19 continued

$\alpha = .05$	MH		AMD		FC	
	F	ES	F	ES	F	ES
			Between Factors			
Ability (D)	314.12***	.31	6.11*	.04	.59	.01
Sample (S)	47.44***	.12	367.15***	.34	258.97***	.28
DIF Size (DS)	297.22***	.31	1980.54***	.79	1405.86***	.66
D × S	26.17***	.09	58.74***	.14	53.00***	.13
D × DS	9.87**	.06	26.96***	.09	.11	.00
S × DS	.72	.01	162.13***	.23	175.47***	.23
D × S × DS	.72	.01	74.88***	.15	59.57***	.14
			Within Factors			
Purif. (P)	112.73***	.19	801.56***	.50	41.97***	.32
D × P	106.47***	.18	23.18***	.09	8.94**	.05
S × P	1.43	.02	33.38***	.10	.11	.00
DS × P	1.81	.02	300.46***	.31	14.60***	.07
D × S × P	.81	.02	.33	.01	.03	.00
D × DS × P	1.81	.02	1.34	.02	2.76	.03
S × DS × P	8.95**	.05	6.27*	.04	5.41*	.04
D × S × DS × P	8.95**	.05	3.71	.03	1.35	.02

*P<.05. **P<.01. ***P<.001.

Table 20

Analysis of Variance Comparing Effects of All Factors on the Detection Rate (Power) for the Logistic Regression Procedure in the Nonsymmetric Nonuniform DIF Datasets

(df=1, 3192)

	LR	
	F	ES
$\alpha = .01$		
Ability Distribution (D)	93.07***	.17
Sample Size (S)	301.55***	.31
DIF Size (DS)	697.47***	.47
D × S	7.04**	.05
D × DS	65.28***	.14
S × DS	257.24***	.28
D × S × DS	1.76	.02
$\alpha = .05$		
Ability Distribution (D)	38.41***	.11
Sample Size (S)	62.11***	.14
DIF Size (DS)	236.53***	.27
D × S	26.28***	.09
D × DS	31.20***	.10
S × DS	52.84***	.13
D × S × DS	20.38***	.08

**P<.01. ***P<.001.

Table 21

Detection Rate (Power) in Each Condition for Each Method in the Nonsymmetric Nonuniform DIF Datasets

	Same Mean				Different Mean			
	Large Sample		Small Sample		Large Sample		Small Sample	
$\alpha = .01$	.80	.40	.80	.40	.80	.40	.80	.40
MH								
No Purification	.50	.27	.50	.21	.94	.53	.67	.32
2-step	.50	.26	.50	.21	.78	.43	.57	.26
Iterative	.50	.26	.50	.21	.90	.52	.67	.32
AMD								
No Purification	1.0	.35	.83	.04	.94	.37	.71	.21
2-step	.99	.42	.82	.04	.95	.37	.72	.20
Iterative	1.0	.87	1.0	.48	1.0	.62	.98	.51
FC								
No Purification	1.0	.57	.91	.09	.98	.43	.82	.25
2-step	1.0	.56	.90	.10	.97	.42	.82	.24
Iterative	1.0	.60	.92	.10	.97	.42	.83	.24
LR	1.0	.96	1.0	.66	1.0	.82	.97	.43

Continued next page

Table 21 continued

	Same Mean				Different Mean			
	Large Sample		Small Sample		Large Sample		Small Sample	
$\alpha = .05$	.80	.40	.80	.40	.80	.40	.80	.40
MH								
No Purification	.51	.36	.50	.27	.98	.70	.85	.49
2-step	.50	.31	.50	.26	.89	.54	.69	.39
Iterative	.51	.31	.50	.26	.95	.67	.78	.45
AMD								
No Purification	1.0	.56	.90	.11	.99	.46	.86	.31
2-step	1.0	.60	.90	.08	.97	.46	.83	.30
Iterative	1.0	.82	.99	.37	1.0	.79	.99	.64
FC								
No Purification	1.0	.72	.97	.32	1.0	.58	.94	.42
2-step	1.0	.80	.97	.29	.99	.60	.94	.45
Iterative	1.0	.82	.98	.31	.99	.67	.97	.49
LR	1.0	.94	1.0	.90	1.0	.93	.99	.72